Acta Scientific NUTRITIONAL HEALTH (ISSN:2582-1423)

Volume 5 Issue 3 March 2021

Research Article

Science/Education Portraits VII: Statistical Methods Used in 1081 Papers Published in Year 2020 Across 12 Life Science Journals Under BioMed Central

Kyle D Kim[1,2], Shaun CH Chua[3] and Maurice HT Ling[1-4]*

[1]*Department of Applied Sciences, Northumbria University, United Kingdom*

[2]*School of Life Sciences, Management Development Institute of Singapore, Singapore*

[3]*School of Applied Sciences, Temasek Polytechnic, Singapore*

[4]*HOHY PTE LTD, Singapore*

***Corresponding Author:** Maurice HT Ling, Department of Applied Sciences, Northumbria University, United Kingdom.

Received: January 22, 2021

Published: February 12, 2021

Abstract

Statistics is an integral part of biology and is required for all undergraduate life science curriculum. However, are biology students trained in statistical skills required in the field? Despite studies listing various commonly statistical methods used in specialised branches of life sciences; such as, immunology and tropical biology; there is a lack of study on the common statistical methods used in life science in general. Here, we examine 1081 articles across 12 life sciences journals under BioMed Central, published in 2020, to elucidate the common statistical methods used in current life science research, as a basis to recommend an updated syllabus to all institutions that educate biologists. 72.7% of the examined articles contains identifiable statistical methods and a total of 2431 instances were identified. Our findings show that the first 3 out of 15 categories of methods; parametric comparison of means (25.38% of instances), correlation/regression (18.88%), and post-hoc test (10.32%); accounts for 54.59% of the instances. In terms of individual methods, the top 8 methods account for 52.04% of the instances – (a) t-test (13.00%), (b) ANOVA (12.26%), (c) unspecified (likely to be Pearson's correlation) and Pearson's correlation (9.79%), (d) Benjamini and Hochberg's False Discovery Rate (FDR) (4.77%), (e) Tukey's HSD (4.36%), (f) Kruskal-Wallis Test (2.96%), (g) Mann-Whitney U Test (2.80%), and (h) Chi Square Test (2.10%). These findings may have an impact on future curriculum design.

Keywords: Statistical Methods; Life Science; Research; Education

Introduction

The importance of statistics in biology has been recognized more than a century ago [1]. With increasing number of statistical methods, there is a concern regarding core statistical fundamentals required in a biologist's education [2-6]. Lee., *et al*. [5] review articles in six pharmacy journals and found (a) ANOVA, (b) Chi-Square Test, (c) Student's t-Test, (d) Pearson's Correlation Coefficient, and (e) Logistic Regression; as the five most commonly used inferential statistical methods. Loaiza Velásquez., *et al*. [2] review the statistical methods used in two tropical journals during a year and identified twelve most frequently used methods as (a) ANOVA, (b) Chi-Square Test, (c) Student's t-Test, (d) Linear Regression, (e) Pearson's Correlation Coefficient, (f) Mann-Whitney U Test, (g) Kruskal-Wallis Test, (h) Shannon's Diversity Index, (i) Tukey's Test,

(j) Cluster Analysis, (k) Spearman's Rank Correlation Test, and (l) Principal Component Analysis. It is crucial that biology students are trained in required statistical skills [2].

However, Lee., *et al*. [5] focus on pharmacy while Loaiza Velásquez., *et al*. [2] focus on tropical biology. Similar work by Skinner [6], Meyr [7], Al-Benna., *et al*. [8], and Hammer and Buffington [9] focus on immunology, surgery, burns research, and veterinary medicine respectively. Hence, the statistical methods used in the common denominator, life science in general, can only be inferred. Here, we identify the statistical methods used in 1081 articles across 12 life sciences journals under BioMed Central published in 2020 as a basis to recommend an updated syllabus to all institutions that educate biologists.

Methods

Using similar methods in previous studies [2,5,7,8], twelve open access journals from BioMed Central that are indexed in PubMed [(a) Biological Research, (b) BMC Bioinformatics, (c) BMC Biology, (d) BMC Ecology, (e) BMC Evolutionary Biology, (f) BMC Genomics, (g) BMC Microbiology, (h) BMC Molecular and Cell Biology, (i) Cell and Bioscience, (j) Genome Biology, (k) Journal of Animal Science and Biotechnology, and (l) Stem Cell Research and Therapy] were selected for survey. For each of the twelve journals, all articles published from January 01, 2020; to the end of the month where the number of articles exceed 100 were chosen, or to the end of October 2020. This is to prevent an over-representation of a specific journal in the survey. For each published article, identifiable statistical method(s) used were collated and each method was recorded only once per article [5] with no judgement made on the suitability of the methods [7].

Results and Discussion

In this study, we examined 1081 peer-reviewed articles published across 12 open access journals from BioMed Central to collate the statistical methods used. The minimum 2-year and 5-year impact factors (as of October 2020) are 2.381 and 2.922 respectively (Table 1), with the highest 2-year impact factor at 10.806; suggesting that the 12 open access journals are highly reputable. Hence, the statistical methods used is likely to be reflective of the needs of the field and important support for curriculum development [2].

Journal Name	2-year Impact Factor	5-year Impact Factor	Source Normalized Impact per Paper (SNIP)	SCImago Journal Rank (SJR)
Biological Research	3.092	2.968	0.939	0.841
BMC Bioinformatics	3.242	3.213	1.156	1.626
BMC Biology	6.765	7.296	1.604	3.698
BMC Ecology	2.381	2.922	0.913	1.030
BMC Evolutionary Biology	3.058	3.252	1.198	1.531
BMC Genomics	3.594	4.093	1.140	1.629
BMC Microbiology	2.989	3.381	1.049	1.154
BMC Molecular and Cell Biology	3.066	2.684	1.023	1.070
Cell and Bioscience	5.026	4.443	0.985	1.410
Genome Biology	10.806	19.041	2.794	9.479
Journal of Animal Science and Biotechnology	4.167	4.392	1.690	1.333
Stem Cell Research and Therapy	5.116	5.554	1.267	1.501

Table 1: Impact Factors and Ranking of Journals (as of October 2020).

Of the 1081 articles examined, 786 (72.7%) articles contain identifiable statistical methods (Table 2). From which, 2431 instances of statistical methods were identified. 51.79% (n = 405; Figure 1) of the articles contain one or two statistical methods; with 14 as the maximum number of statistical methods identified from a single article [Tran., *et al*. [10]]. The methods identified were categorized into 15 application categories (Table 3). The top 3 categories; (a) parametric comparison of means, (b) correlation/regression, and (c) post-hoc test; accounts for 54.59% of the instances. These are followed by (a) non-parametric comparison of

means, (b) multiple comparison correction, (c) dimension reduction/multidimensional scaling, and (d) goodness of fit test; which accounts for another 26.94% of the instances. Collectively, these 7 application categories accounts for 81.53% of the instances.

Journal Name	Date Range	Number of Articles Surveyed	Number of Articles with Statistical Methods
Biological Research	January 01, 2020 to November 30, 2020	55	48 (87.3%)
BMC Bioinformatics	January 01, 2020 to February 28, 2020	72	31 (43.1%)
BMC Biology	January 01, 2020 to June 30, 2020	111	79 (71.2%)
BMC Ecology	January 01, 2020 to December 31, 2020	69	55 (79.7%)
BMC Evolutionary Biology	January 01, 2020 to July 31, 2020	96	51 (53.13%)
BMC Genomics	January 01, 2020 to January 31, 2020	109	68 (62.39%)
BMC Microbiology	January 01, 2020 to April 30, 2020	99	81 (81.0%)
BMC Molecular and Cell Biology	January 01, 2020 to November 30, 2020	85	71 (83.5%)
Cell and Bioscience	January 01, 2020 to September 30, 2020	106	54 (50.9%)
Genome Biology	January 01, 2020 to April 30, 2020	100	83 (83.0%)
Journal of Animal Science and Biotechnology	January 01, 2020 to September 30, 2020	97	91 (93.8%)
Stem Cell Research and Therapy	January 01, 2020 to February 28, 2020	82	74 (90.2%)
Total		1081	786 (72.7%)

Table 2: Number of Articles Surveyed.

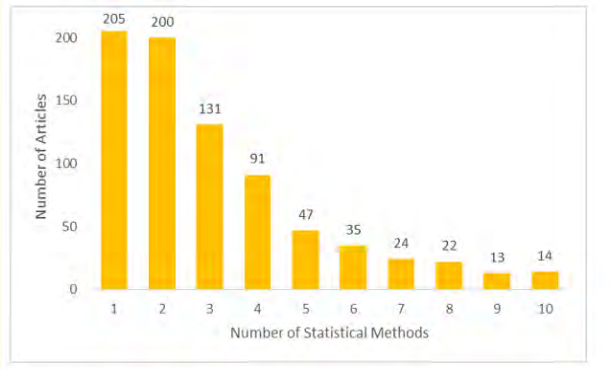

Figure 1: Distribution of Number of Statistical Methods.

In terms of individual statistical methods, the top 8 most frequent methods account for 52.04% of all the instances (Table 4). The methods are (a) t-test (13.00%), (b) ANOVA (12.26%), (c) unspecified and Pearson's correlation (9.79%), (d) Benjamini and Hochberg's False Discovery Rate (FDR) (4.77%), (e) Tukey's HSD (4.36%), (f) Kruskal-Wallis Test (2.96%), (g) Mann-Whitney U Test (2.80%), and (h) Chi Square Test (2.10%). These results are consistent with that of Loaiza Velásquez., *et al*. [2] as 7 of the 8 methods are common, except FDR. These results are also generally consistent with the common statistical methods identified by Al-Benna., *et al*. [8] and Meyr [7].

Mann-Whitney U Test is often used as the non-parametric equivalent of independent samples t-test in most cases [11] despite differences in several assumptions [12]. Kruskal-Wallis Test is essentially the non-parametric version of one-way ANOVA [13]. While this underpins the importance of parametric and non-parametric comparison of means to biological sciences as these 4 methods account for 31.02% of the instances, it also illustrates the importance of non-parametric methods in biological sciences as biological data is often not normally distributed [14-16]. Besides being not normally distributed, multiple testing is also common in biology [17,18]; hence, it is not surprising that FDR is a commonly seen in publications.

Chi Square test is a common statistical method in biology with applications from clinical sciences [19] to population genetics [20] to omics analyses [21,22]. It is also often the first statistical test

taught in first year genetics; hence, has an important position in biology. Similar to Chi Square test, correlation is a staple in many fields of biology [23,24]. Of the 12 major post-hoc tests, Tukey's HSD is the only one that appears in the top 8 most frequent methods. One of the reasons may be its simplicity and closeness to t-test assuming equal variance in terms of calculation [25]. This also demonstrates the importance of t-test in the education of a biologist as 3 of the 8 most frequent methods (Tukey's HSD, ANOVA, and Mann-Whitney U Test) requires pre-requisite knowledge of t-test. The presence of Tukey's HSD also suggests the need for biologists to know what to do after null hypothesis of equal means in more than 2 samples, such as in ANOVA, is rejected. Taken together, these 8 statistical methods should form the basis of all statistical curriculum for biologists.

Application Category	Frequency	Prevalence (%)	Cumulative Prevalence (%)
Parametric Comparison of Means	617	25.38%	25.38%
Correlation/ Regression	459	18.88%	44.26%
Post-Hoc Test	251	10.32%	54.59%
Non-parametric Comparison of Means	231	9.50%	64.09%
Multiple Comparison Correction	167	6.87%	70.96%
Dimension Reduction/ Multidimensional Scaling	133	5.47%	76.43%
Goodness of Fit Test	124	5.10%	81.53%
Graphing	101	4.15%	85.68%
Normality Test	59	2.43%	88.11%
Omics Analysis	49	2.02%	90.13%
Randomization/ Permutation Test	36	1.48%	91.61%
Survival Analysis	34	1.40%	93.01%
Equality of Variance test	32	1.32%	94.32%
Measure of Dispersion	13	0.53%	94.86%
Others	125	5.14%	100.00%

Table 3: Relative Prevalence of Statistical Methods. Prevalence is defined as the quotient between the number of frequencies in each category and the total number of frequencies (n = 2431). Cumulative prevalence is the summation of prevalence up to the category, with the total cumulative prevalence of 100%

Application Category/ Statistical Method	Frequency (N)	Prevalence (%)
Correlation/Regression (18.88%)		
Unspecified Correlation	147	6.05%
Pearson's correlation	91	3.74%
Spearman's correlation	46	1.89%
General Linear Model (GLM)	33	1.36%
Linear Regression	28	1.15%
Unspecified Regression	15	0.62%
Linear Mixed Effect Model	10	0.41%
Polynomial Contrasts	9	0.37%
ADONIS	7	0.29%
Cox's Regression	6	0.25%
Generalized Linear Mixed Model	6	0.25%
Logistic Regression	6	0.25%
Non-Linear Regression	6	0.25%
Mantel-Haenszel Method	5	0.21%
Multiple Regression	5	0.21%
Inter-Rater Agreement	4	0.16%
Kendall's Correlation	4	0.16%
Redundancy Analysis (RDA)	3	0.12%
Fractional Regression	3	0.12%
Generalized Additive Model	3	0.12%
Lasso Regression	3	0.12%
Others (N < 3)	19	0.78%
Dimension Reduction/Multidimensional Scaling (5.47%)		
Principal Component Analysis (PCA)	40	1.65%
Principal Co-Ordinates Analysis (PCoA)	17	0.70%
Linear Discriminant Analysis (LDA)	15	0.62%
Linear Discriminant Analysis Effect Size (LEfSe)	14	0.58%
Shannon Index	10	0.41%
Simpson Index	6	0.25%
Neighbor-Joining Method	5	0.21%
Non-Metric Multidimensional Scaling (NMDS)	4	0.16%
Root Mean Squared Distance (RMSD)	4	0.16%
Uniform Manifold Approximation and Projection	4	0.16%
UPGMA Cluster Analysis	4	0.16%

t-Distributed Stochastic Neighbor Embedding	3	0.12%
Others (N < 3)	7	0.29%
Equality of Variance Test (1.32%)		
Levene's Test	13	0.53%
F-Test	9	0.37%
Bartlett's Test	5	0.21%
Others (N < 3)	5	0.21%
Goodness of Fit Test (5.10%)		
Chi Square Test	51	2.10%
Fisher's Exact Test	48	1.97%
Likelihood Ratio Test	6	0.25%
Mantel Test	5	0.21%
Wald Test	5	0.21%
Hardy-Weinberg Equilibrium Test	4	0.16%
I2 Test	3	0.12%
Kullback–Leibler Distance	2	0.08%
Graphing (4.15%)		
Box Plot	29	1.19%
Heatmap	22	0.90%
Error Bar	12	0.49%
Bar Plot	9	0.37%
Scatter Plot	7	0.29%
Dot Plot	4	0.16%
Receiver Operating Characteristic Curve (ROC curve)	4	0.16%
Volcano Plot	4	0.16%
Others (N < 3)	10	0.41%
Measure of Dispersion (0.53%)		
Coefficient of Variation	6	0.25%
Standard Error	4	0.16%
Others (N < 3)	3	0.12%
Multiple Comparison Correction (6.87%)		
Benjamini and Hochberg's False Discovery Rate (FDR)	116	4.77%
Bonferroni Correction	48	1.97%
Others (N < 3)	3	0.12%
Non-Parametric Comparison of Means (9.50%)		
Kruskal-Wallis Test	72	2.96%
Mann-Whitney U Test	68	2.80%
Wilcoxon Rank-Sum Test	40	1.65%
Analysis of Similarities (ANOSIM)	22	0.90%
Unspecified Wilcoxon Test	12	0.49%
Wilcoxon Signed-Rank Test	11	0.45%
Friedman Test	3	0.12%
Others (N < 3)	3	0.12%
Normality Test (2.43%)		
Shapiro-Wilk Test	27	1.11%
Kolmogorov-Smirnov Test	25	1.03%
Anderson-Darling Test	3	0.12%
Others (N < 3)	4	0.16%
Omics Analysis (2.02%)		
Gene Ontology Enrichment Analysis	24	0.99%
Analysis of Molecular Variance (AMOVA)	6	0.25%
DESeq2	3	0.12%
Feed Conversion Ratio	3	0.12%
Others (N < 3)	13	0.53%
Parametric Comparison of Means (25.38%)		
t-Test	316	13.00%
Analysis of Variance (ANOVA)	298	12.26%
Z-Test	3	0.12%
Post-Hoc Test (10.32%)		
Tukey's HSD	106	4.36%
Unspecified Post-Hoc Test	44	1.81%
Duncan's multiple range Test	23	0.95%
Dunnett's Test	20	0.82%
Dunn's Test	17	0.70%
Fisher's Least Significant Difference (LSD)	13	0.53%
Holm-Sidak Test	6	0.25%
Student–Newman–Keuls (SNK) Test	5	0.21%
Scheffe's Test	4	0.16%
Bonferroni Test	3	0.12%
Conover-Iman Test	3	0.12%
Post-Hoc t-Test	3	0.12%
Others (N < 3)	4	0.16%
Randomization/Permutation Test (1.48%)		
Permutational Multivariate Analysis of Variance (PERMANOVA)	21	0.86%
Permutation Test	7	0.29%
Others (N < 3)	8	0.33%
Survival Analysis (1.40%)		
Log-Rank Test	24	0.99%
Kaplan-Meier Analysis	10	0.41%
Others (N < 3)	125	5.14%

Table 4: Breakdown of Statistical Methods.

Prevalence is defined as the quotient between the number of frequencies in each category and the total number of frequencies (n = 2431). Cumulative prevalence is the summation of prevalence up to the category, with the total cumulative prevalence of 100%.

Conclusion

The top 8 most frequent methods identified from 1081 articles are (a) t-test, (b) ANOVA, (c) unspecified and Pearson's correlation, (d) Benjamini and Hochberg's False Discovery Rate (FDR), (e) Tukey's HSD, (f) Kruskal-Wallis Test, (g) Mann-Whitney U Test, and (h) Chi Square Test.

Supplementary Materials

Data file for this study can be downloaded at http://bit.ly/SEP7_Statistics.

Conflict of Interest

The authors declare no conflict of interest.

Disclaimer

The views expressed by the authors are that of the authors rather than the views of their affiliated institutions.

Bibliography

1. Pearl R. "The Service and Importance of Statistics to Biology". *American Statistical Association* 14.105 (1914): 40-48.
2. Loaiza Velásquez N., *et al*. "Which Statistics Should Tropical Biologists Learn?" *Revista de Biología Tropical* 59.3 (2011): 983-992.
3. Colon-Berlingeri M and Burrowes PA. "Teaching Biology Through Statistics: Application of Statistical Methods in Genetics and Zoology Courses". *CBE—Life Sciences Education* 10.3 (2011): 259-267.
4. Metz AM. "Teaching Statistics in Biology: Using Inquiry-Based Learning to Strengthen Understanding of Statistical Analysis in Biology Laboratory Courses". *CBE—Life Sciences Education* 7.3 (2008): 317-326.
5. Lee CM., *et al*. "Statistics in the Pharmacy Literature". *Annals of Pharmacotherapy* 38.9 (2004): 1412-1418.
6. Skinner J. "Statistics for Immunologists". *Current Protocols in Immunology* 122.1 (2018): 54.
7. Meyr AJ. "A 5-Year Review of Statistical Methods Presented in The Journal of Foot and Ankle Surgery". *The Journal of Foot and Ankle Surgery* 49.5 (2010): 471-474.
8. Al-Benna S., *et al*. "Descriptive and Inferential Statistical Methods Used in Burns Research". *Burns* 36.3 (2010): 343-346.
9. Hammer AS and Buffington CA. "Survey of Statistical Methods Used in the Veterinary Medical Literature". *Journal of the American Veterinary Medical Association* 205.2 (1994): 344-345.
10. Tran HTN., *et al*. "A Benchmark of Batch-Effect Correction Methods for Single-Cell RNA Sequencing Data". *Genome Biology* 21.1 (2020): 12.
11. Dowdy S., *et al*. "Statistics for Research". John Wiley and Sons (2011).
12. Fay MP and Proschan MA. "Wilcoxon-Mann-Whitney or t-test? On Assumptions for Hypothesis Tests and Multiple Interpretations of Decision Rules". *Statistics Surveys* 4 (2010): 1-39.
13. Kruskal WH and Wallis WA. "Use of Ranks in One-Criterion Variance Analysis". *Journal of the American Statistical Association* 47.260 (1952): 583-621.
14. Mar JC. "The Rise of the Distributions: Why Non-Normality is Important for Understanding the Transcriptome and Beyond". *Biophysical Reviews* 11.1 (2019): 89-94.
15. Bono R., *et al*. "Non-normal Distributions Commonly Used in Health, Education, and Social Sciences: A Systematic Review". *Frontiers in Psychology* 8 (2017): 1602.
16. Wittkowski KM and Song T. "Nonparametric Methods for Molecular Biology". *Methods In Molecular Medicine* 620 (2010): 105-153.
17. Korthauer K., *et al*. "A Practical Guide to Methods Controlling False Discoveries in Computational Biology". *Genome Biology* 20.1 (2019): 118.
18. Saxon E. "Multiple Comparisons". *BMC Biology* 13.1 (2015): 86.
19. Al-Najjar D., *et al*. "CoVID-19 Symptoms Analysis of Deceased and Recovered Cases using Chi-square test". *European Review for Medical and Pharmacological Sciences* 24.21 (2020): 11428-11431.

20. Kamarudin NJ., *et al*. "A Simulation Study on the Effects of Founding Population Size and Number of Alleles Per Locus on the Observed Population". *EC Veterinary Science* 5.8 (2020): 176-180.

21. Li W-M., *et al*. "Prognostic Utility of FBLN2 Expression in Patients With Urothelial Carcinoma". *Frontiers in Oncology* 10 (2020): 570340.

22. Wang C., *et al*. "High Expression of RING Finger Protein 126 Predicts Unfavorable Prognosis of Epithelial Ovarian Cancer". *Medical Science Monitor* 26 (2020): e921370.

23. Mukaka MM. "Statistics Corner: A Guide to Appropriate Use of Correlation Coefficient in Medical Research". *Malawi Medical Journal* 24.3 (2012): 69-71.

24. Armstrong RA., *et al*. "The Use of Correlation and Regression Methods in Optometry". *Clinical and Experimental Optometry* 88.2 (2005): 81-88.

25. Haynes W. "Tukey's Test". In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, editors. Encyclopedia of Systems Biology. New York, NY: Springer New York (2013): 2303-2304.